DOCUMENT RESUME

ED 206 646 TM 810 535

AUTHOR Lockwood, Robert E.
TITLE A Procedure to Select and Rank Qualified Applicants Using Perceived Item Difficulty and Perceived Item Relevance.
PUB DATE Apr 81
NOTE 16p.: Paper presented at the Annual Meeting of the National Council on Measurement in Education (Los Angeles, CA, April 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores: *Difficulty Level: Employment Potential: Job Skills: Minimum Competency Testing: *Occupational Tests: *Personnel Selection: *Scoring Formulas: Test Validity
IDENTIFIERS Angoff Methods: Lawshe (C): *Ranking: Uniform Guidelines Employee Selection Procedures

ABSTRACT

To replace traditional methods of minimum competency testing in personnel selection, a two stage cutting score procedure is proposed. The Angoff procedure is modified so that only a subset of test items, which assess essential job related skills, are used to determine the cutting score and to identify qualifed applicants. As a measure of item content validity the Lawshe rating scale is used by a panel of judges. Essential items are identified, and difficulty indices estimated for a hypothetical population of minimally qualified examinees: the sum of the item probabilities, over judges, equals the minimum cutting score. The second procedural stage involves the identification of test items assessing useful, but not essential skills. The number of these items answered correctly, plus the number of correct essential items is used to rank qualified candidates. This procedure provides a rational methodology for producing an applicant eligibility list for jobs requiring specific skills and abilities in view of constraints imposed by the Uniform Guidelines on Employee Selection Procedures. (AEF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED206646

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A PROCEDURE TO SELECT AND RANK QUALIFIED APPLICANTS USING PERCEIVED ITEM DIFFICULTY AND PERCEIVED ITEM RELEVANCE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. E. Lockwood

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ROBERT E. LOCKWOOD

EDUCATIONAL TESTING SERVICE

ATLANTA, GEORGIA

TM 810 535

Prepared for presentation as part of the symposium "Standard Setting: Procedures, Practices, and Constraints." at the Annual Meeting of the National Council on Measurement in Education, Los Angeles, April 1981.

A PROCEDURE TO SELECT AND RANK QUALIFIED APPLICANTS
USING PERCEIVED ITEM DIFFICULTY AND PERCEIVED ITEM RELEVANCE

The practice of setting cutting scores, and, indeed, the whole philosophy underlying that effort, takes on additional complications when any cutting score methodology is moved from classroom and/or certification-type examinations to the realm of personnel selection examinations. The cutting score methodologies most frequently used in certification testing require judges to establish a minimal standard below which an examinee will be considered incompetent. In personnel selection examinations the focus is totally reversed, and the examination "consumer" is less interested in identifying those few examinees who cannot do the job than he/she is interested in selecting, from among a list of qualified applicants, the most qualified, the smaller number of persons to actually fill available positions.

Secondly, traditional cutting score methodologies are usually freed from the necessity of addressing ranking considerations such as those introduced by the Uniform Guidelines on Employee Selection Procedures, and to which all merit system practices must pay heed.

To amplify, most cutting score methodologies use the opinions of subject matter experts as the basis for establishing the level of minimal competence. For example, in a written test situation these experts estimate the probability--for an assumed "minimally competent" examinee--of a correct response to each item. The sum of these various probabilities is then totaled for all items and all judges and an average computed, this average representing the cutting score for the test. Underlying such

practices is, of course, the assumption that each test item measures a knowledge, skill or ability essential to the area tested. In other words, which particular items are answered is of less consequence than the mere fact that the required number of items are answered.

However, for ranking purposes in situations where one has a large applicant pool and a minimal number of available positions, these assumptions require re-thinking. That is, it seems logically contradictory to assume in one case that examinee performance will allow the identification of those examinees who will be unsuccessful (i.e., cannot do a job or cannot be trained to do the job) and to assume in a second case (but using the same methodology) that the same items and the examinees' performance on them will allow one to draw fine distinctions among qualified examinees whose performance exceeded the minimal standard. To simply rank order candidates on the basis of their total test scores introduces a conflict of assumptions. In setting the minimum score, it has been assumed that all items measure an essential skill, knowledge or ability. Thus in theory a perfect score should be the minimum criterion. Some allowance can be made for chance errors, but it is not consistent with an assumption of essentiality to then discriminate, using chance errors, among candidates, all of whom have been judged to have all essential skills.

In recognition of this dilemma and the specific needs of examiners working under personnel and merit system constraints, the following two-stage cutting score procedure is proposed. This procedure will take as its major premise the assumption that not all items on the selection test measure minimally essential skills.

The first stage of the two-stage process involves setting the minimum acceptable level of test performance for successful job performance. Many procedures have been proposed that make use of expert judgment to determine this standard (e.g., Nedelsky, 1954; Angoff, 1971; Ebel, 1972). One such procedure proposed by Angoff asks "...each judge to state the probability that the 'minimally acceptable person' would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score." (Angoff, 1971, p. 515). Two changes are, however, proposed to the traditional way in which Angoff's method is used. First, it appears reasonable to explain to the judges the concept of the "chance score" (i.e., the score likely to result from mere random guessing). In a multiple-choice test, such as that used to exemplify the approach described in this paper, this score is simply the inverse of the number of options. That is, a 4-choice item would yield a chance score of 1/4 or .25. The importance of the chance score for our discussion and for our judges is, of course, that it has the effect of restricting the range of probabilities assigned to an item or the estimate of the number of minimally acceptable persons who would answer the item correctly.

The second change to Angoff's procedure reflects a concern over which of the test items are included in the summation process to determine the final cutting score. In most cutting score procedures, all items in the test are used to determine the cutting score. However, the proposed procedure uses only a subset of the total test items, this subset defined via a technique proposed by Lawshe (1975). This technique was proposed

initially as a mechanism for indexing both the content validity of the items on a test and the test as a whole. Further, the index, a quantification of content validity, is derived from the judgments of job experts concerning the essentiality of the skill or knowledge being measured to the performance of a critical job task identified through some job analytic procedure. To arrive at the quantification, each judge is asked to rate each item on the scale shown below.

"Is the skill (or knowledge) measured by this item

_____ essential,

_____ useful but not essential, or

_____ not necessary

to the performance of the task?"

Responses from all judges are pooled once the number indicating "essential" for each item is determined. The quantification or content validity ratio (CVR) is defined as

$$CVR = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

where n_e equals the number of judges indicating that the skill tapped by the item was essential.

This procedure yields a value ranging from -1 to +1, with a CVR = 0 indicating a situation where the judges were evenly split in their estimates of essentiality. Therefore, an item with a CVR greater than zero has some degree of content validity and, as the CVR approaches 1, the greater can be said to be the degree of content validity. Lawshe (1975) provides the following statement concerning the validity of the judgments.

> Whenever panelists or other experts make judgments, the question properly arises as to the validity of their judgments. If the panelists do not agree regarding the essentiality of the knowledge or skill measured to the performance of the job, then serious questions can be raised. If, on the other hand, they do agree, we must conclude that they are either "all wrong" or "all right." Because they are performing the job, or are engaged in the direct supervision of those performing the job, there is no basis upon which to refute a strong consensus. (p. 567)

In the present paper, it is the CVR that will be used to select the subset of items which determine the cutting score. That is, only those items which yielded CVR values reflecting agreed-upon essentiality will be included within this critical subset. Such a step is proposed since there appears to be a logical argument for excluding, from the pool used to set a "minimum" standard, those items which assess skills or knowledges not essential to job performance. Thus, items regarded as non-essential or items whose essentiality indices might reasonably have occurred through chance will be eliminated from consideration during the first step of the procedure. The following table was used to indicate the minimum value of the CVR.

TABLE 1
Minimum Values of CVR
One Tailed Test, p = .10

No. of Panelists	CVR	No. of Panelists to Agree
5	.76	4.4
6	.70	5.1
7	.63	5.7
8	.58	6.3
9	.53	6.9
10	.50	7.5
11	.47	8.1
12	.45	8.7
13	.43	9.3
14	.41	9.9
15	.40	10.5
20	.34	13.4
25	.30	16.2
30	.27	19.1
40	.23	24.6

___ When the number saying "essential" is more than half, but less than all, the CVR is somewhere between zero and 1.00.

To summarize the essence of the proposed melding of the Lawshe and Angoff methodologies, the first stage of the proposed procedure involves identifying those items that assess skills or knowledges judged essential to job performance and the estimation for each of those items, difficulty indices for a population of hypothetical minimally qualified examinees. The sum of the item probabilities, over judges, will be the minimum cutting score.

The second stage of the process involves a method for logically ranking those applicants who score above the minimum score required on the "essential" items. This is necessary for the following reason. In situations where a single test is used to determine and then rank qualified applicants, the ranking supposes two conflicting conditions exist. First, that all items measure an essential skill or knowledge and second, that these items can rank order examinees based on the percentage of essential items answered correctly. The truth of the matter, however, is that performance on essential items can be used to determine qualified vs. not qualified examinees, but not to rank order qualified candidates. To rank order qualified candidates requires that the variance in test performance be related to differences in job performance. Specifically, the Uniform Guidelines on Employee Selection Procedures (1978) Section 4C (9) state:

> Ranking based on content validity studies. If a user can show, by a job analysis or otherwise, that a higher score on content valid selection procedure is likely to result in better job performance, the results may be used to rank persons who score above minimum levels. Where a selection procedure supported solely or primarily by content validity is used to rank job candidates, the selection procedure should measure those aspects of performance which differentiate among levels of job performance.

To demonstrate that a higher score is likely to result in better job performance, a different procedure must be employed. To rank order applicants on test performance demands that the items measure skills or abilities that differentiate levels of performance among qualified persons. Recalling the procedure proposed by Lawshe, the category of "useful but not essential" clearly indicates those items that would differentiate levels of performance. Therefore, the second stage of the procedure, this procedure intended to

result in the ranking of applicants, is simply the identification of those items that assess useful but not essential skills or knowledges for task performance, and the summation of those items that were answered correctly by a qualified examinee. ("Qualified" being defined as having scored above the minimum acceptable level on the subset of essential items identified in the first stage of the procedure.) Again it is proposed that a content validity ratio be established for these items using the formula

$$CVR_2 = \frac{n_u - \frac{N}{2}}{\frac{N}{2}}$$

where n_u equals the number of judges who rated the item as "either (a) essential or (b) useful, but not essential" and that Table 1 be used to determine the critical values of the CVR_2. At this stage the first two categories from Lawshe's scale were collapsed to determine those items which might be used for ranking purposes. This was performed since it logically follows that if half the judges rate an item essential and half rate the item useful but not essential, we are assured that all agree to its usefulness but not to its essentiality. Usefulness is a necessary but not sufficient condition for essentiality and it is the useful items that must be identified.

Those items that are included in this stage would be scored and the number correct, combined with the score for those items rated as essential, would be used to rank the qualified applicants.

Thus, for each examinee, there would be two scores, a score on the essential items and a score for the total test (i.e., a combined score

representing performance on both items rated essential and items rated as useful but not essential). The first subset score would be compared to the established cutting score. If the applicant score is below the cutting score, the applicant is classified unqualified; if the score is above the cutoff, the applicant is classified qualified. For the qualified applicants, their scores on the total of "essential " and "useful" items would be used to rank order them for hiring priority.

The test used in piloting the procedure described above was a 90-item 4-choice multiple-choice instrument intended to be used for selecting employees for several state social service agencies. Specifically, applicants' scores would be used to rank those applicants on a hiring eligibility list which would, in turn, be made available to a diverse array of agencies having job vacancies. These agencies had participated in the cross-agency job analysis procedure which served as the genesis of the test in question, and thus the tasks represented by the test encompassed behaviors relevant to all of the positions studied.

In recognition of the way in which the eligibility list resulting from the test scores would be used, it was determined that the panel of judges to be involved in the experimental procedure should represent, to the extent possible, the agencies involved. Accordingly, ten incumbents were released by their employer agencies to serve as the panel reviewing the examination. The size of this panel, combined with the diversity of positions represented, yielded certain methodological difficulties. (These will be dealt with at a later point in the paper.)

The specific steps required of the panel of judges in completing the experimental procedure are as follows. First, panel members estimated the item difficulty for each item on the test, using as a point of reference a hypothetical minimally competent examinee (Angoff, 1971). This step was performed initially--rather than at a later point--in an attempt to control for any biasing effect which might result were judges to have first rated the items using the essentiality scales described earlier. One departure from the Angoff procedure as typically described was the discussion, with the judges, of the impact random guessing has on the probability of correctly answering an item. This modification was intended to restrict the range of probabilities assigned by the judges (i.e., from .00 to 1.00 as is typically allowed in the "classical" Angoff procedure to the more realistic .25 to 1.00). Prior to actual implementation of the procedure group practice and discussion was held to assure that the judges held a common concept of minimal competency.

Second, panel members estimated the relevance of each item to task performance using the essentiality scales proposed by Lawshe (1975). It should be noted that the scales proposed by Lawshe were modified slightly. Specifically, the essentiality of the skill or knowledge measured by the item was related to the specific task performance measured by the item and not the more general realm of overall job performance. This modification was employed since the examination in question was developed using a task based job analysis, as opposed to a procedure emphasizing knowledges, skills and abilities (KSA's), and because the examination was intended for use by multiple agencies for multiple jobs. As was done with the Angoff procedure, actual implementation of the Lawshe procedure was preceded by a period of discussion and practice.

The data obtained from the judges were treated in the following manner. First data obtained using the Lawshe procedure were analyzed to identify the population of items regarded as essential. This population of items would serve as the focus for the Angoff type cutting score procedure.

The reader will recall that mention was made earlier in this paper of certain methodological problems resulting from the size/composition of the panel of judges. Specifically these difficulties involved the computation of the CVR critical to Lawshe's procedure. Table 1 presents the minimum value, given the number of judges comprising the panel which will allow the researcher to conclude with 90% confidence that more than 50% of the judges concur regarding the essentiality of the item. For a panel of 10 judges as was used in the present study, such a CVR necessitates agreement by 7.5 judges. Thus two options exist, the first is the more conservative approach and would require agreement by 8 of the 10 judges. This would result in a confidence level of 94%. For the present study this would result in the identification of ten items as essential for performance. The second option was to be less conservative and more pragmatic and require agreement by only 7 of the 10 judges. This results in a confidence level of 83%. Because of the aforementioned diversity of the judges it was decided to take the less conservative approach for the current study. Thus, the required CVR was determined to be .40 rather than the .50 indicated in Table 1. Using this value, 37 of the 90 items were identified as being essential for task performance.

Using the 37 items thus identified the researcher summed the item difficulties determined by the panel of judges using the Angoff procedure. The resultant summed figure, 32, thus represents the cutting score used

in determining a pool of qualified examinees (i.e., any examinee earning less than a raw score of 32 on the 37 items identified as essential would be relegated to the pool of unqualified examinees and would be dropped from any position on the eligibility list). Remembering that the criterion for qualification which might be used to define a qualified examiner would be perfect performance (i.e., 37 out of 37) the cutting score of 32 represents a recognition of the previously mentioned chance error that might occur on essential items.

What is then required of the procedure is a means to rank the qualified examiners (i.e., those who score 32 or more on the 37 essential items). To do this, one component would certainly be the examinees' score on those items that the panel of judges indicated as measuring useful, but not essential, aspects of task performance. These items assess the aspects of performance required by the Uniform Guidelines on Employee Selection Procedures for the ranking of examinees. This study identified 53 items as being useful for job performance (using the collapsed categories mentioned earlier) and thus appropriate for rank ordering applicants. However, it would appear to be shortsighted for any ranking procedure to ignoree differences in examinee performance on those items deemed essential to task performance (i.e., the 37 items which formed the basis for the initial cutting score). Thus, the pool of examinees initially identified as being qualified may be rank ordered for hiring eligibility using the combination of the useful but not essential items and the essential items. This not only uses all available data on the applicants but also serves to maintain the reliability of the examination.

It should be mentioned that in most situations items regarded as not essential will have been dropped from the test form during an earlier developmental review stage. However, situations may exist in which an agency may, for pragmatic reasons, wish to use an examination developed externally to

screen applicants for positions which are highly similar--though not identical--to the position represented by the agency wishing to use the test. In such situations it is anticipated that the panel of judges will, during the Lawshe procedure described, eliminate all inappropriate items for the position in question. Thus such items will not be included in the pool of items used to screen examinees into those who are qualified/not qualified nor will performance on these items contribute toward the ranking of the pool of applicants identified as qualified.

This procedure provides a methodology for creating an eligibility list that is both rational and defensible in terms of the constraints placed on employing agencies by the Uniform Guidelines or Employee Selection Procedures. However, it is not proposed that this procedure is a panacea for all the ills of employee selection but rather a device that can be used under certain employment conditions (another tool in the testers tool box as it were). It would seem that these conditions would include a degree of control over the hiring of applicants from the eligibility list, situations where the applicant pool is larger than the available number of positions and possibly when there exists a greater degree of homogenity of the tasks comprising the job than was present in this pilot study. These situations may correspond to the first row of the second two columns of Dr. Temp's matrix found on page 12 of his paper (Temp, Goodrich, Weber, & Lockwood, 1981). That is, situations where the occupation is of medium high risk and the applicant pool is greater than available positions. This methodology would seem less than useful where there exists multiple diverse jobs or multiple agencies using the same eligibility list for hiring; situations where attitudinal factors may influence job performance/retention; and in situations where other factors (e.g., availability to work in a specific location) exert influence on hiring priority.

REFERENCES

Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike (Ed.) Educational Measurement (2nd. ed.). Washington, D.C.: American Council on Education, 1971.

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Lawshe, C. H. A quantitative approach to content validity. Personnel Psychology, 1975, 28, 563-575.

Temp, G. E., Goodrich, M., Weber, M. B., & Lockwood, R. E. Issues associated with setting standards in Civil Service selection systems. In J. E. McLean (Chair), Standard setting: Procedures, practices, and constraints. Symposium presented at the meeting of the National Council on Measurement in Education, Los Angeles, 1981.

Uniform guidelines on employee selection procedures. Federal Register, August 25, 1978, Part IV.